Data Analysis and Visualization in R (IN2339)

Exercise Session 6 - Graphically supported hypotheses

Daniela Klaproth-Andrade, Felix Brechtmann, Julien Gagneur

Section 00 - Getting ready

1. Make sure you have already installed and loaded the following libraries:

```
library(ggplot2)
library(data.table)
library(magrittr)   # Needed for %>% operator
library(tidyr)
```

Section 01 - Color guidelines

What are best practices when using color for data visualizations? Select all that apply.

1. Avoid having too many colors for categorical data.
2. Use one bright color to attract the readers attention.
3. Use color only when it actually adds meaning to the plot.
4. Use divergent color scales for categorical data types.

```
# Correct are 1, 3
```

Section 02 - Confounding factors

Investigate the file `coffee_sim.csv` by first loading it as a `data.table`.

```
coffee_dt <- fread("./extdata/coffee_sim.csv")
coffee_dt
```

```
##       V1      risk packs_per_day cups_per_day
##   1:   0  3.514369             0          1-5
##   2:   1  6.338370             0          1-5
##   3:   2  2.173321             0          1-5
##   4:   3  4.152559             0           5+
##   5:   4  6.091390             0          1-5
##  ---
## 196: 195 15.033181            2+            0
## 197: 196 14.692839            2+            0
## 198: 197 15.759205            2+            0
## 199: 198 17.125810            2+           5+
## 200: 199 14.378317            2+          1-5
```

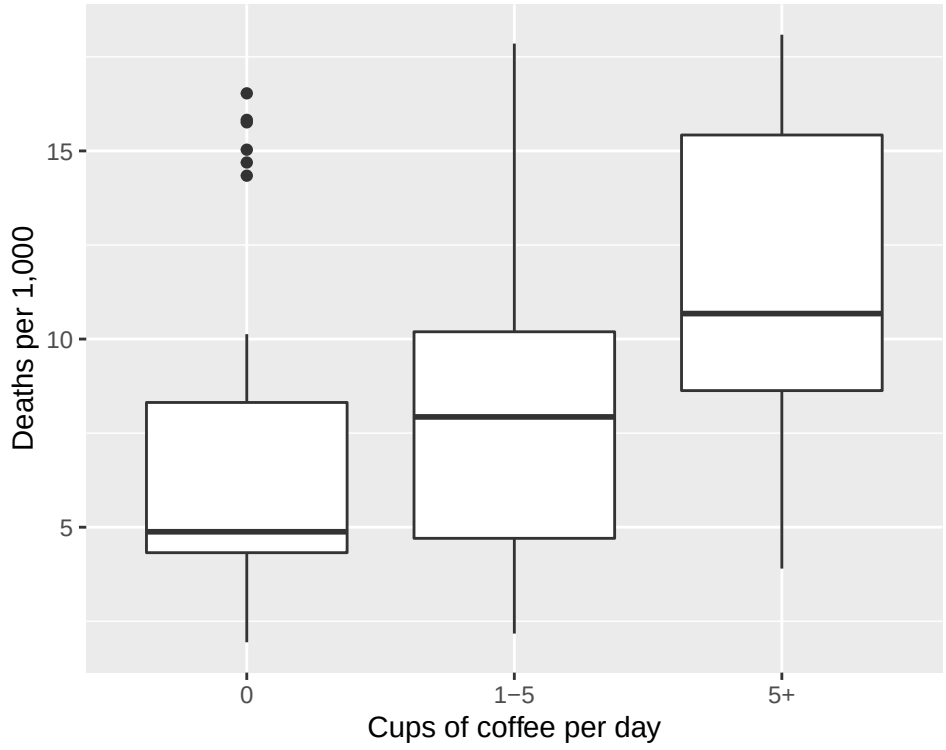
```
summary(coffee_dt)
```

```
##        V1              risk        packs_per_day      cups_per_day
##  Min.   :  0.00   Min.   : 1.943   Length:200         Length:200
##  1st Qu.: 49.75   1st Qu.: 4.700   Class :character   Class :character
```

```
##  Median : 99.50   Median : 7.258   Mode  :character   Mode  :character
##  Mean   : 99.50   Mean   : 8.182
##  3rd Qu.:149.25   3rd Qu.:10.218
##  Max.   :199.00   Max.   :18.088
```

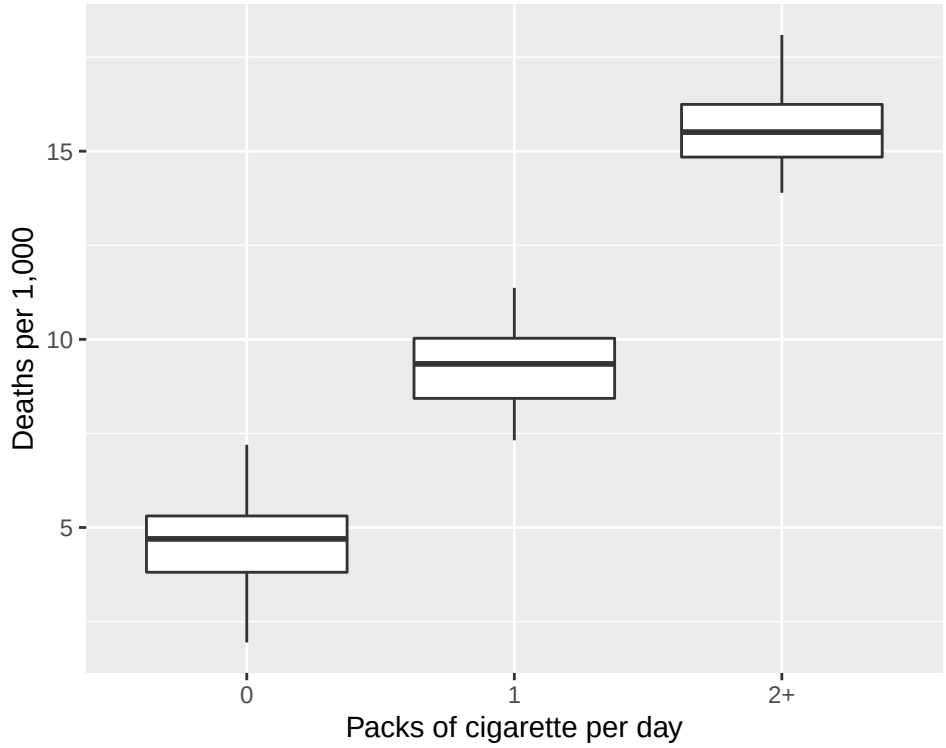
1. Visualize the trend between coffee and coronary heart disease (CHD)-related deaths (risk), which suggests a possible causal relationship.

```
# Confounded association of coffee
ggplot(coffee_dt, aes(cups_per_day, risk)) +
  geom_boxplot() +
  labs(x = "Cups of coffee per day",
       y = "Deaths per 1,000")
```

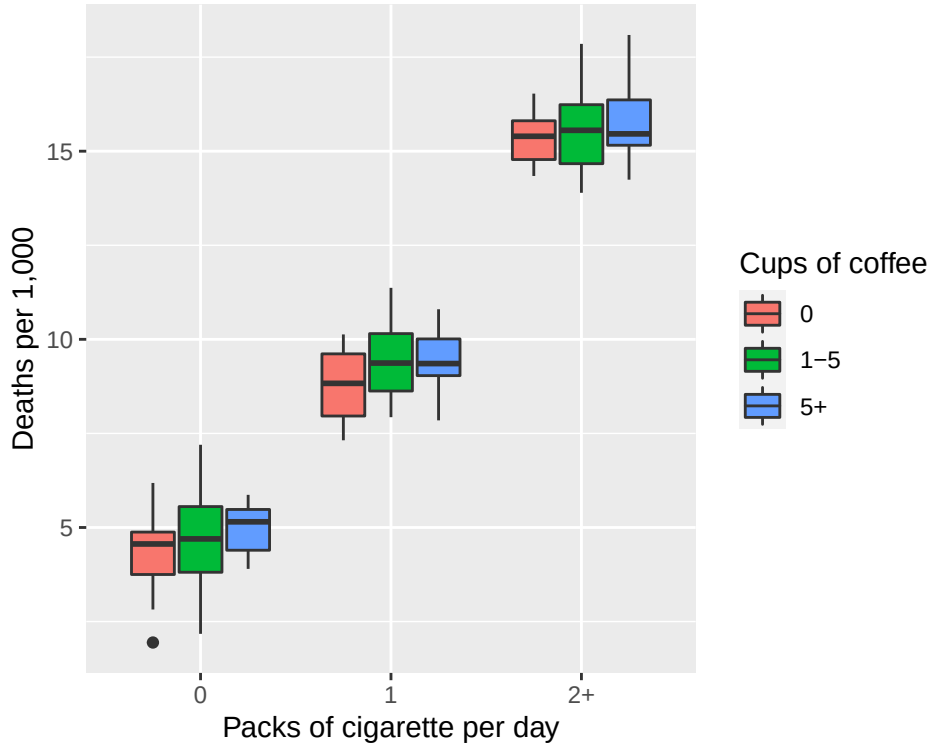


2. From this plot you could conclude that coffee causes CHD. Do you think this conclusion explains the original observation? Provide plots supporting other conclusions.

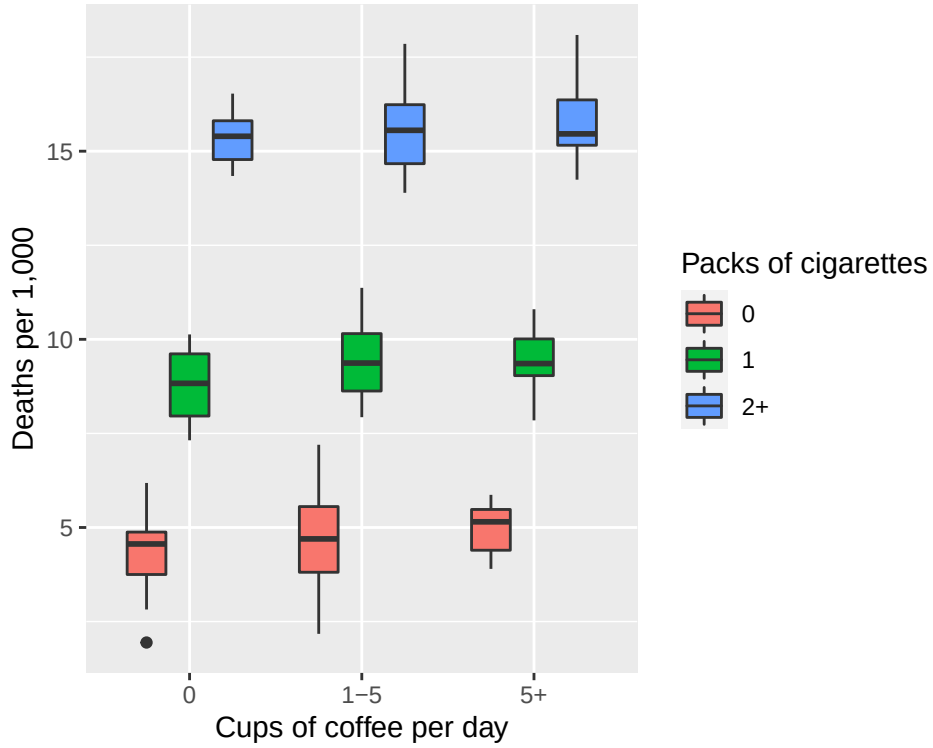
```
# This is the way it looks for smoking
ggplot(coffee_dt, aes(packs_per_day, risk)) +
  geom_boxplot() +
  labs(x = "Packs of cigarette per day",
       y = "Deaths per 1,000")
```




```
# and this is the proper way to look at it,
# coffee effects are always the same within each smoking group.
ggplot(coffee_dt, aes(packs_per_day, risk, fill = cups_per_day)) +
  geom_boxplot() +
  labs(x = "Packs of cigarette per day",
       y = "Deaths per 1,000") +
  guides(fill = guide_legend(title = "Cups of coffee"))
```




```
# But the effect of smoking is not the same within each
# coffee consumption group.
ggplot(coffee_dt, aes(cups_per_day, risk, fill = packs_per_day)) +
  geom_boxplot() +
  labs(x = "Cups of coffee per day",
       y = "Deaths per 1,000") +
  guides(fill = guide_legend(title = "Packs of cigarettes"))
```



Section 03 - Supporting hypotheses with visualizations

1. Read the `titanic.csv` file into a `data.table`. You can read the description of the dataset on kaggle: https://www.kaggle.com/c/titanic/data.

```
##' Load data
titanic <- fread("./extdata/titanic.csv")
titanic
```

```
##       pclass survived                                             name    sex
##    1:      1        1                    Allen, Miss. Elisabeth Walton female
##    2:      1        1                   Allison, Master. Hudson Trevor   male
##    3:      1        0                     Allison, Miss. Helen Loraine female
##    4:      1        0             Allison, Mr. Hudson Joshua Creighton   male
##    5:      1        0 Allison, Mrs. Hudson J C (Bessie Waldo Daniels) female
##   ---
## 1305:      3        0                           Zabour, Miss. Hileni female
## 1306:      3        0                          Zabour, Miss. Thamine female
## 1307:      3        0                      Zakarian, Mr. Mapriededer   male
## 1308:      3        0                            Zakarian, Mr. Ortin   male
## 1309:      3        0                             Zimmerman, Mr. Leo   male
##         age sibsp parch ticket     fare   cabin embarked boat body
##    1: 29.00     0     0  24160 211.3375      B5        S    2   NA
##    2:  0.92     1     2 113781 151.5500 C22 C26        S   11   NA
##    3:  2.00     1     2 113781 151.5500 C22 C26        S        NA
##    4: 30.00     1     2 113781 151.5500 C22 C26        S       135
##    5: 25.00     1     2 113781 151.5500 C22 C26        S        NA
##   ---
## 1305: 14.50     1     0   2665  14.4542               C       328
## 1306:    NA     1     0   2665  14.4542               C        NA
```

```
## 1307: 26.50     0     0   2656   7.2250               C       304
## 1308: 27.00     0     0   2670   7.2250               C        NA
## 1309: 29.00     0     0 315082   7.8750               S        NA
##                            home.dest
##    1:                     St Louis, MO
##    2: Montreal, PQ / Chesterville, ON
##    3: Montreal, PQ / Chesterville, ON
##    4: Montreal, PQ / Chesterville, ON
##    5: Montreal, PQ / Chesterville, ON
##   ---
## 1305:
## 1306:
## 1307:
## 1308:
## 1309:
```

2. Describe what you see in the data. Have a look at the first and last observations. Make a summary of the variables in the dataset.

```
summary(titanic)
```

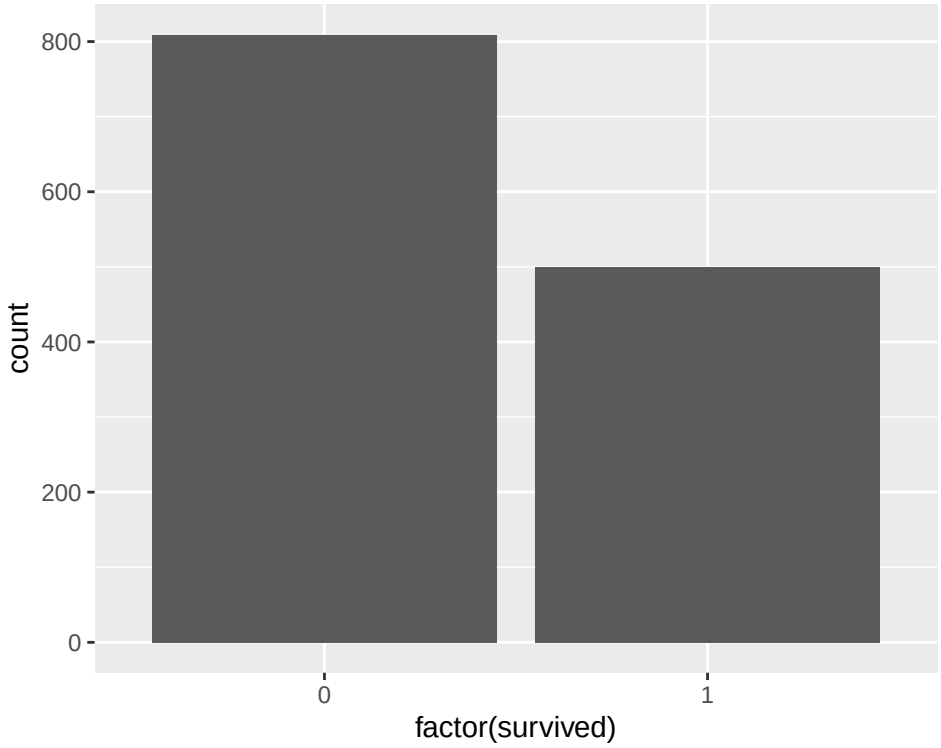
```
##      pclass          survived         name                sex
##  Min.   :1.000   Min.   :0.000   Length:1309        Length:1309
##  1st Qu.:2.000   1st Qu.:0.000   Class :character   Class :character
##  Median :3.000   Median :0.000   Mode  :character   Mode  :character
##  Mean   :2.295   Mean   :0.382
##  3rd Qu.:3.000   3rd Qu.:1.000
##  Max.   :3.000   Max.   :1.000
##
##       age             sibsp            parch          ticket
##  Min.   : 0.17   Min.   :0.0000   Min.   :0.000   Length:1309
##  1st Qu.:21.00   1st Qu.:0.0000   1st Qu.:0.000   Class :character
##  Median :28.00   Median :0.0000   Median :0.000   Mode  :character
##  Mean   :29.88   Mean   :0.4989   Mean   :0.385
##  3rd Qu.:39.00   3rd Qu.:1.0000   3rd Qu.:0.000
##  Max.   :80.00   Max.   :8.0000   Max.   :9.000
##  NA's   :263
##       fare            cabin              embarked            boat
##  Min.   :  0.000   Length:1309        Length:1309        Length:1309
##  1st Qu.:  7.896   Class :character   Class :character   Class :character
##  Median : 14.454   Mode  :character   Mode  :character   Mode  :character
##  Mean   : 33.295
##  3rd Qu.: 31.275
##  Max.   :512.329
##  NA's   :1
##       body        home.dest
##  Min.   :  1.0   Length:1309
##  1st Qu.: 72.0   Class :character
##  Median :155.0   Mode  :character
##  Mean   :160.8
##  3rd Qu.:256.0
##  Max.   :328.0
##  NA's   :1188
```

```
titanic[, table(survived)]
```

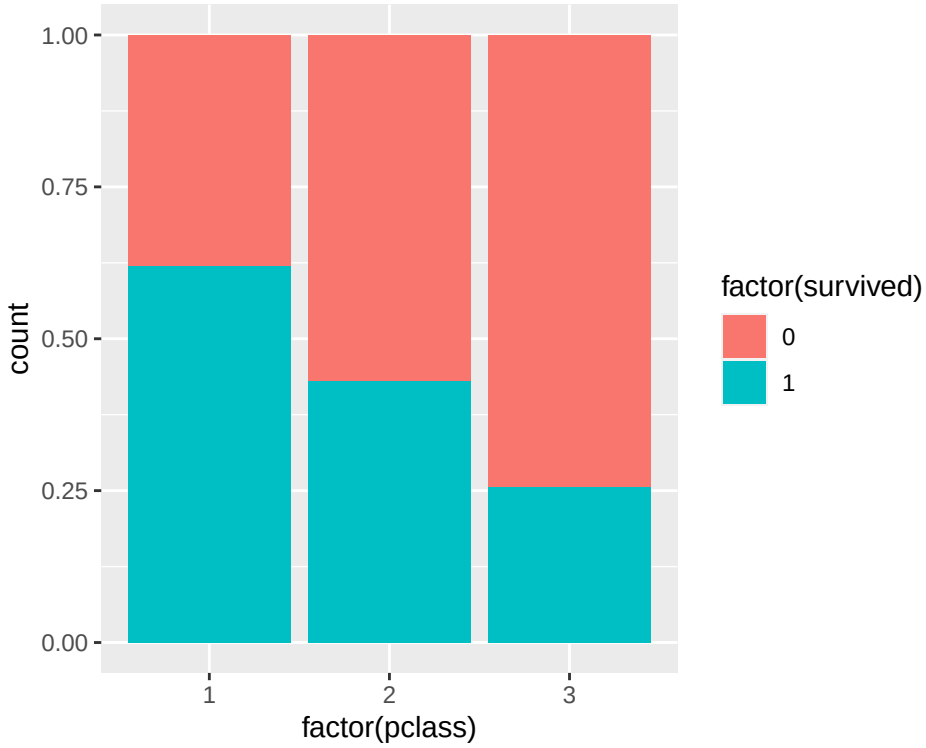
```
## survived
##   0   1
## 809 500
```

3. What do you think are the factors that have the strongest influence on the survival rate? Make claims and justify your argument with plots. *Hint*: check variables like `pclass`, `sex` and `age`, and visualize whether they associate with survival. Additionally check their interactions.

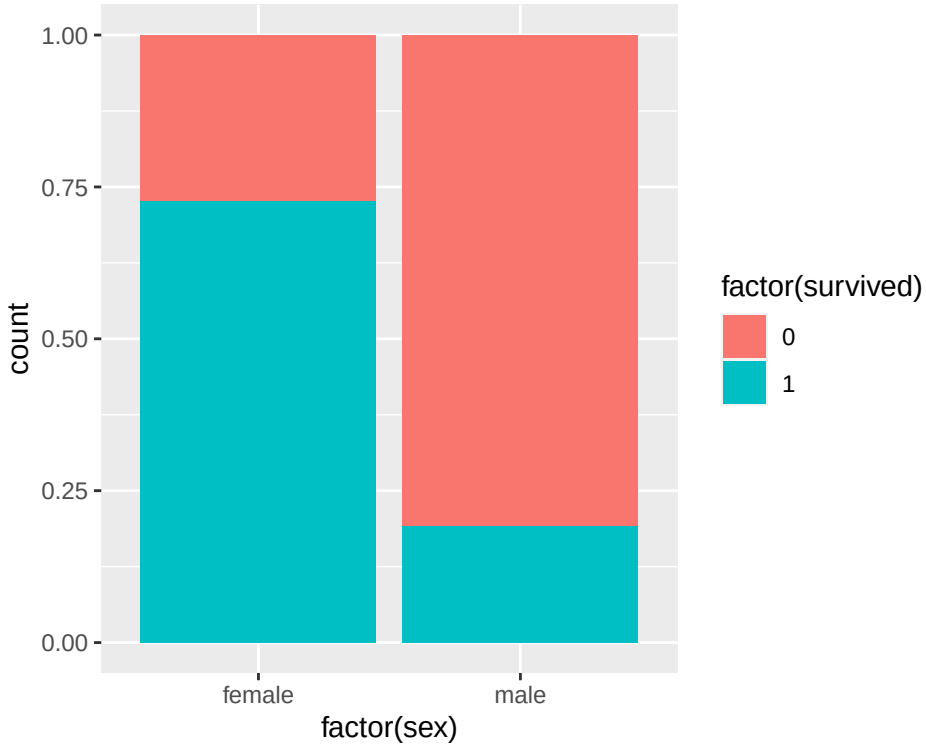
```
# How many survived?
ggplot(titanic, aes(factor(survived))) +
  geom_bar()
```



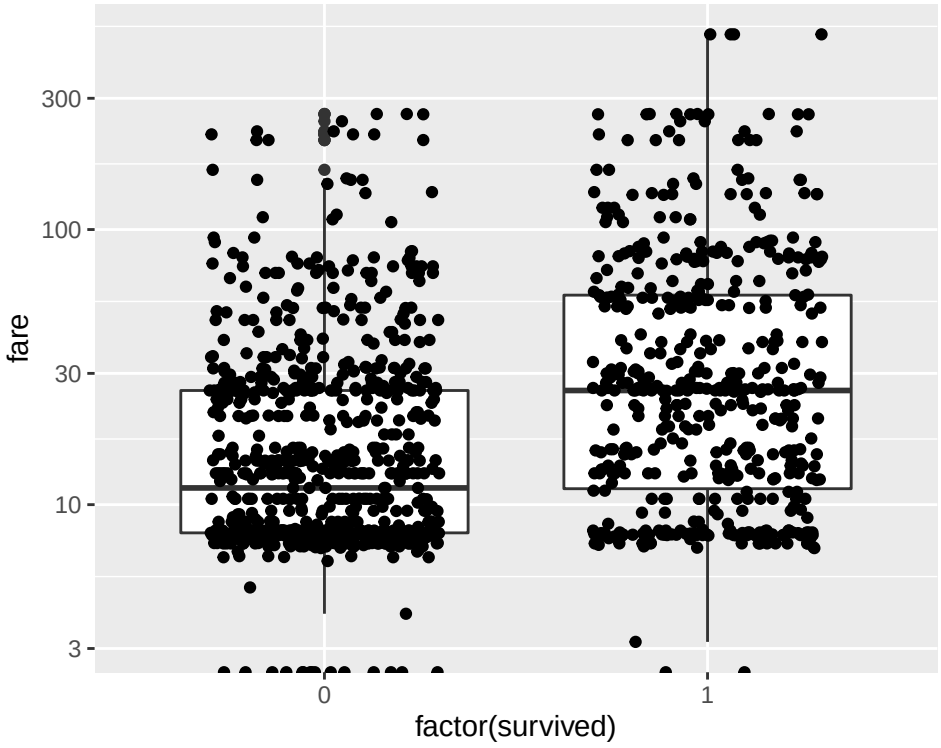

```
# Does passenger class play a role?

# We can see below that the better the class, the higher is the survival rate.
ggplot(titanic, aes(x = factor(pclass), fill = factor(survived))) +
  geom_bar(position = 'fill')
```



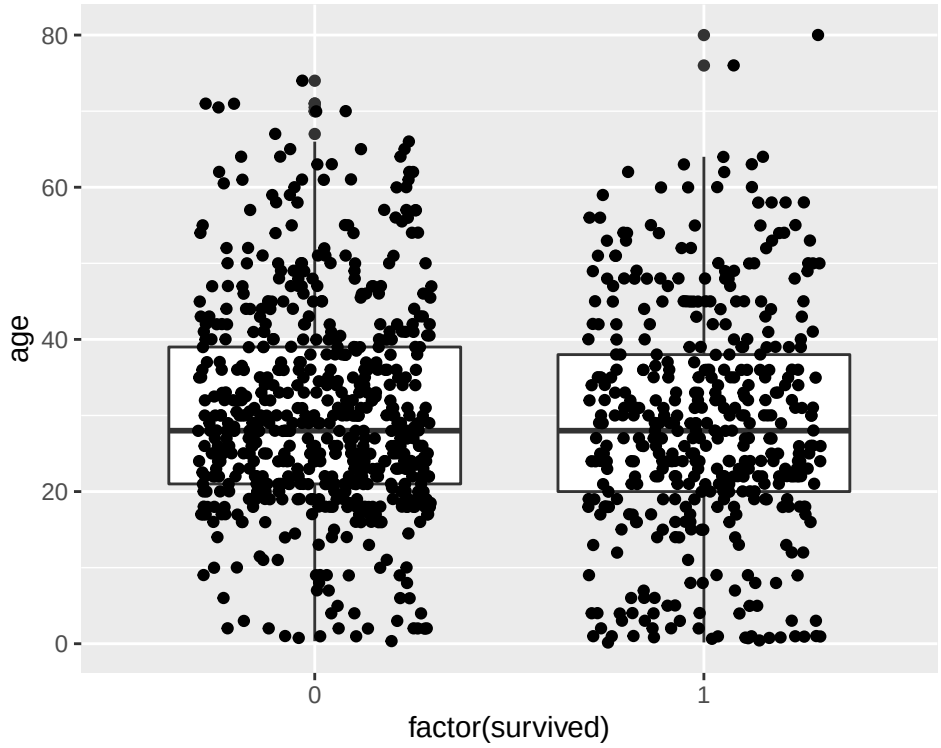

```
# Does sex play a role?
ggplot(titanic, aes(factor(sex), fill = factor(survived))) +
  geom_bar(position = 'fill')
```



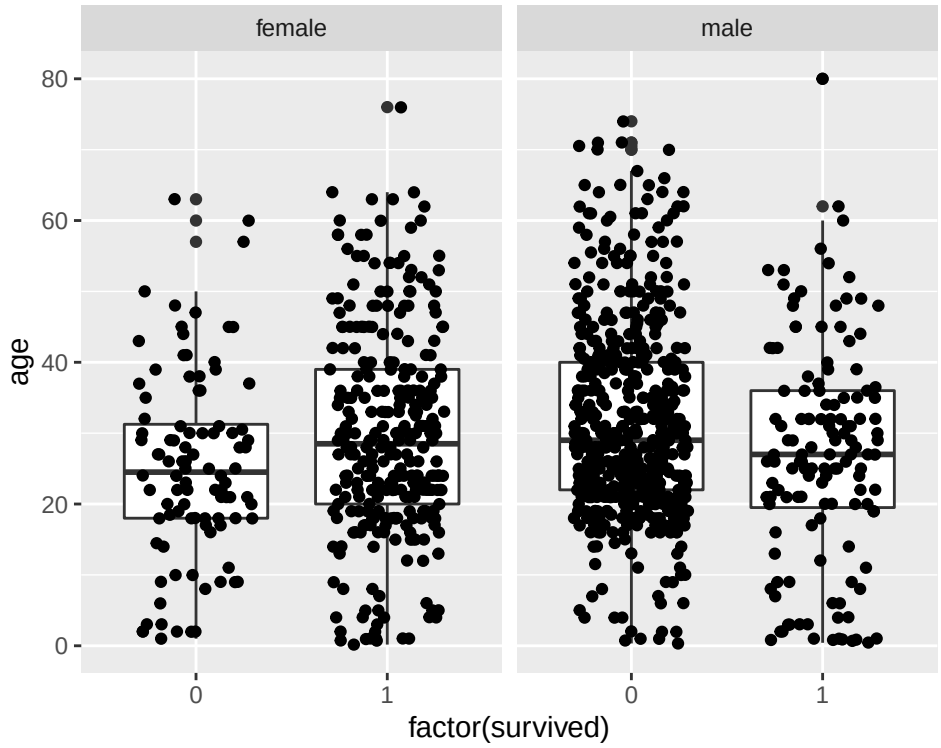
```r
# Does the fare paid play a role?
ggplot(titanic, aes(factor(survived), fare)) +
  geom_boxplot() +
  geom_jitter(width = 0.3) +
  scale_y_log10()
```



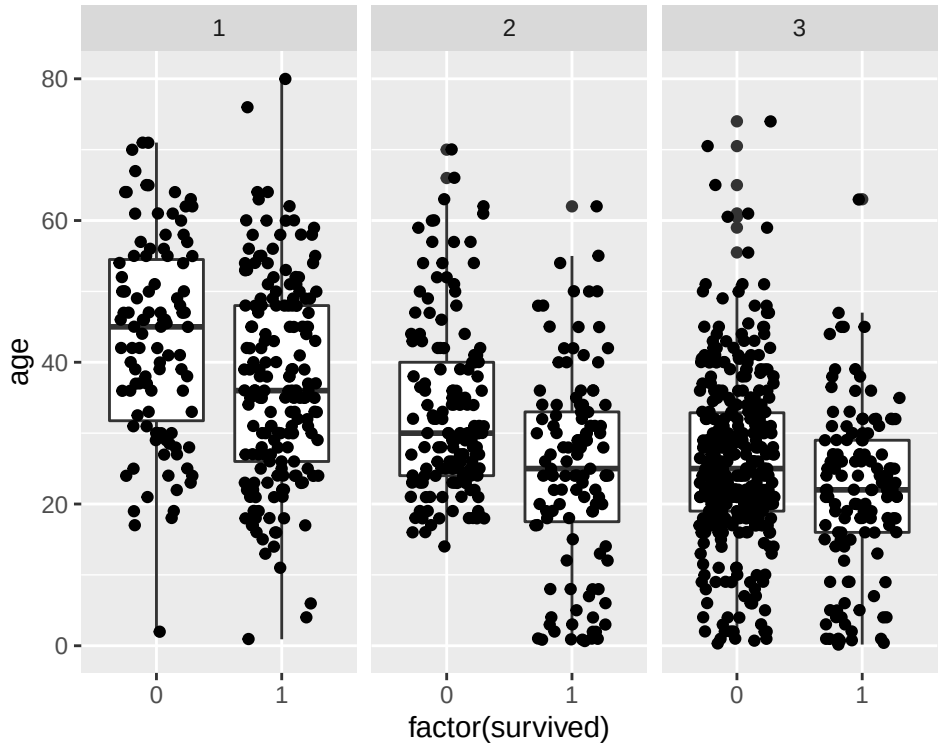

```r
# Does age play a role?

# It seems that age does not associate with survival.
ggplot(titanic, aes(factor(survived), age)) +
  geom_boxplot() +
  geom_jitter(width = 0.3)
```



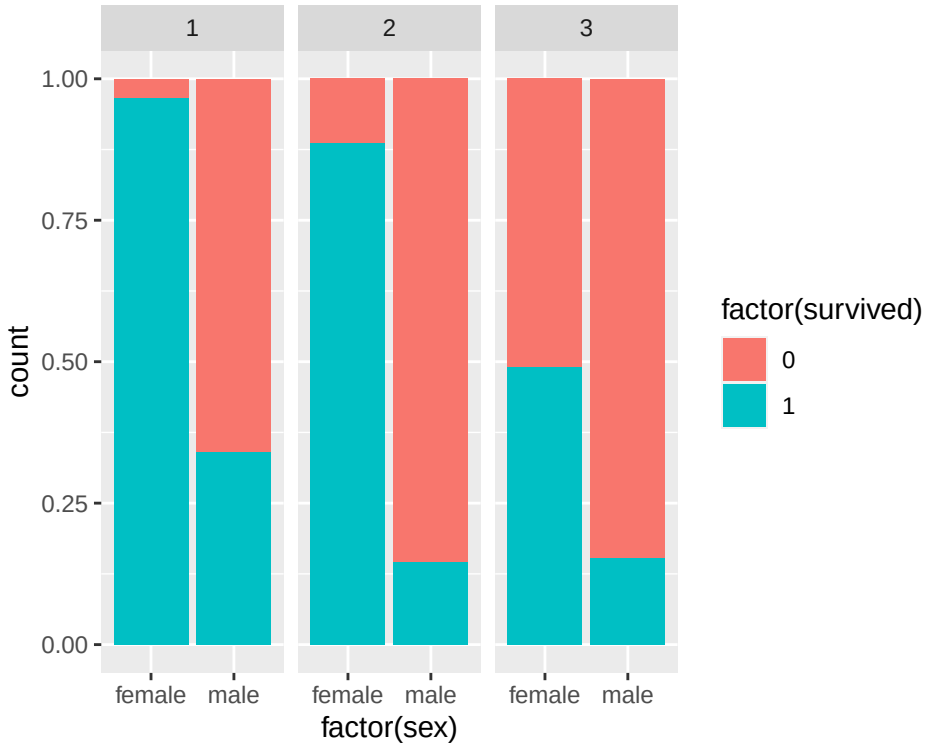

```
# Does this mean age plays almost no role?
# We could investigate the influence of age when we control for sex or the class.
ggplot(titanic, aes(factor(survived), age)) +
  geom_boxplot() +
  geom_jitter(width = 0.3) +
  facet_wrap(~ sex)
```



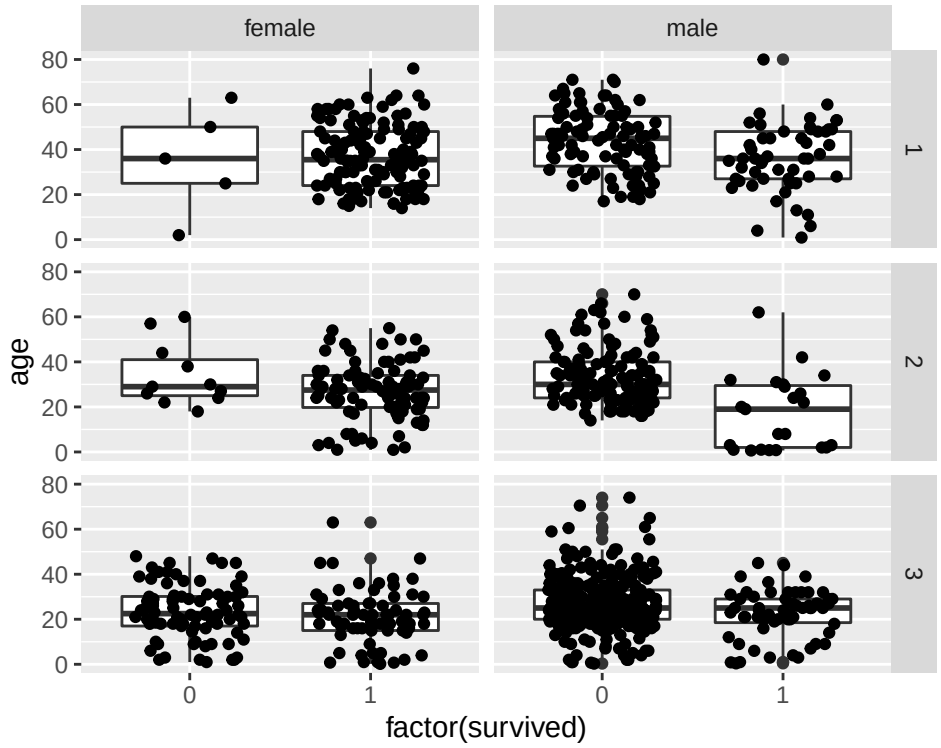

```
# Below we can observe that within each class being younger increased the chances of surviving.
ggplot(titanic, aes(factor(survived), age)) +
  geom_boxplot() +
  geom_jitter(width = 0.3) +
  facet_wrap(~ pclass)
```



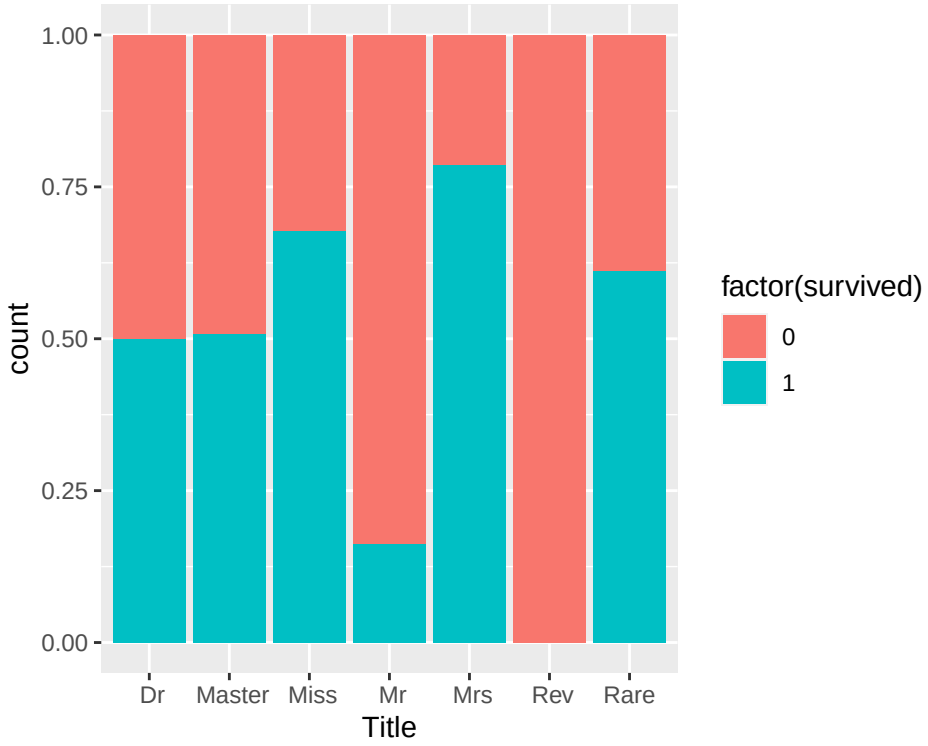

```
# We can additionally check the interaction between gender and the passenger class or control for both

# Interaction between gender and pclass
ggplot(titanic, aes(factor(sex), fill = factor(survived))) +
  geom_bar(position = 'fill') +
  facet_wrap(~ pclass)
```




```
# Controlling for both

# Below we can observe that the gender had huge impact in the first two classes.
# Here a much higher fraction of women than men survied.
# Additionally we can observe that the claim we made above that being younger
# increased the chances of surviving is mostly true for men.
ggplot(titanic, aes(factor(survived), age)) +
  geom_boxplot() +
  geom_jitter(width = 0.3) +
  facet_grid(pclass ~ sex)
```




```
# [OPTIONAL] Check if the title has an impact on the survival.
titanic <- separate(titanic, name, into = c("Name_A", "Name_B"), sep = ",")
titanic <- separate(titanic, Name_B, into = c("Title", "Name"), sep = ". ")
rare_title <- names(table(titanic[,Title])[table(titanic[,Title]) < 7])
titanic$Title[titanic$Title %in% rare_title] <- "Rare"
ggplot(titanic, aes(Title, fill = factor(survived))) +
  geom_bar(position = 'fill')
```



Section 04 - General guidelines in data visualization

Below is a graph taken from one published paper. Read the figure legend.

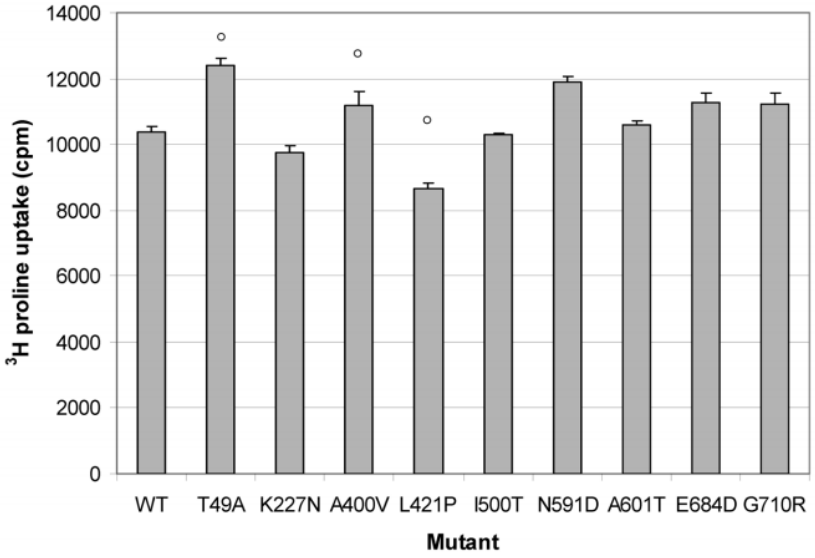


Figure 2. Maximal ^{3}H proline uptake of wildtype (WT) and all tested mutants. The maximum in uptake was measured in the presence of 3 µM cold L-proline. Data are expressed as means ± standard deviation (SD) obtained from triplicate samples. Mutants with a circle were tested in a second independent experiment.
doi:10.1371/journal.pone.0068645.g002

1. [OPTIONAL] Discuss good and bad graphical properties of the plot, make suggestions on how to improve.

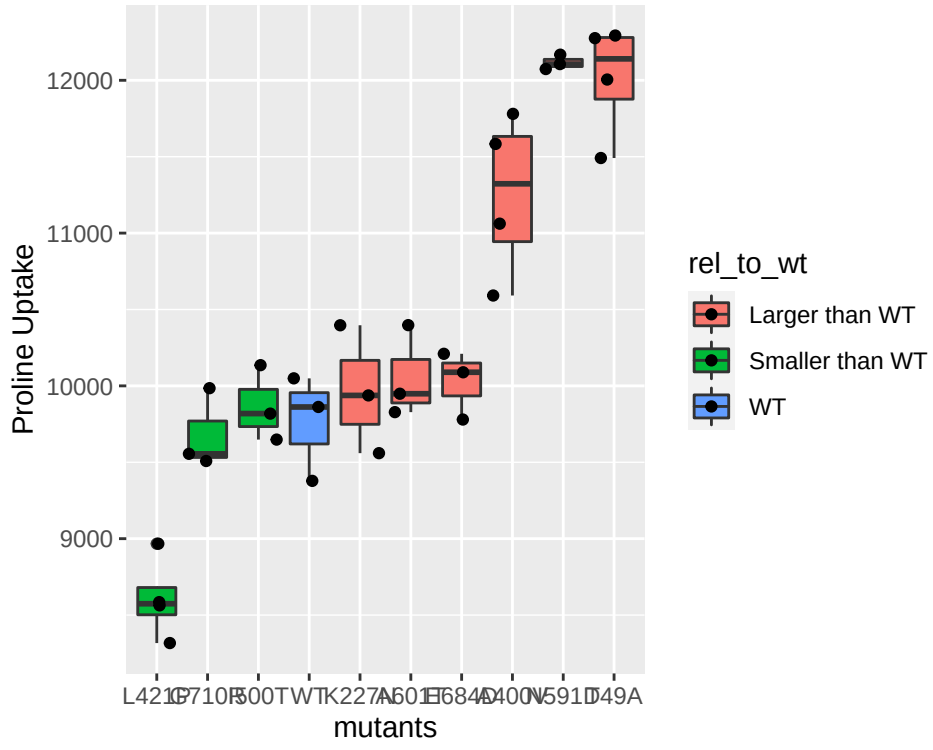
```
# GOOD
# - simple design
# - not too many colors
# - clear labels
# - no chart junk
#
# BAD
# - no highlight, e.g. by color
# - x-axis not sorted
# - summary by mean+sd hides the data, which is at most four points per bar
#
# Suggestion
# - plot single points instead of bars, with small median line (too few points for boxplot)
# - sort Mutants by median
# - give color for above and below WT
```

2. [OPTIONAL] Implement a better visualization. As the original data is not available, we use the data

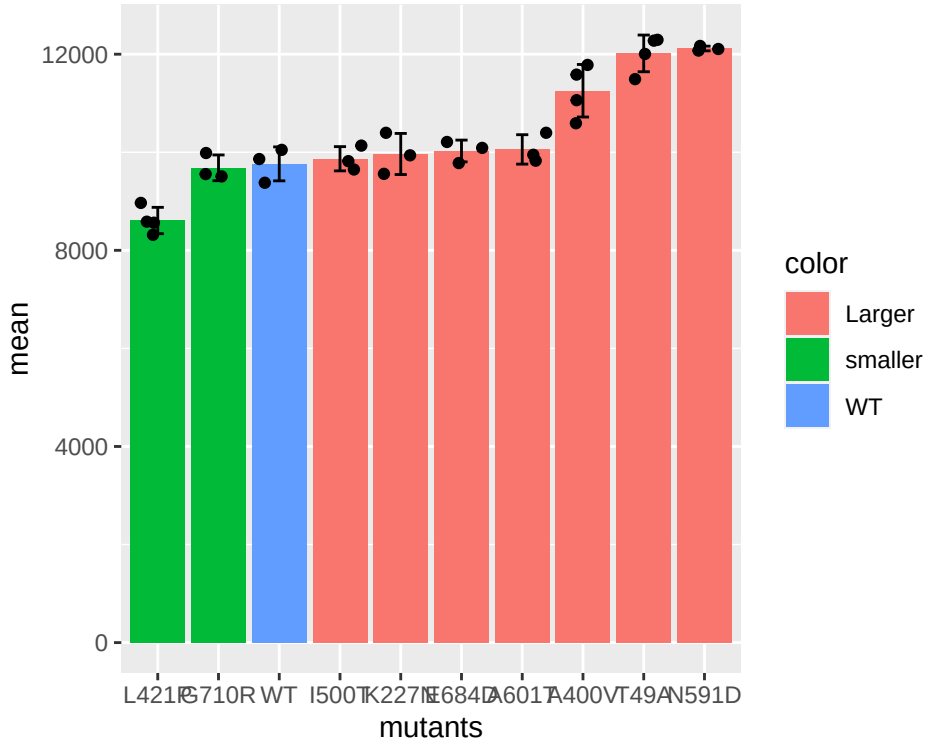
simulated with the code below.

```
# simulate data
dt <- data.table(pro_uptake = c(rnorm(3, 10100, 300), rnorm(4, 12100, 300),
                                rnorm(3, 9850, 300), rnorm(4, 11100, 300),
                                rnorm(4,8300, 300), rnorm(3,10050, 300),
                                rnorm(3, 12000, 300), rnorm(3, 10020, 300),
                                rnorm(3, 10080, 300), rnorm(3, 10070, 300)  ),
                 mutants = c(rep('WT',3), rep('T49A',4), rep('K227N',3), rep('A400V',4),
                             rep('L421P',4), rep('I500T',3), rep('N591D',3),
                             rep('A601T',3), rep('E684D',3), rep('G710R',3) )          )
```

```
# sort by median
dt[, median_per_mut := median(pro_uptake), by = mutants]
wt_med = unique(dt[mutants == 'WT', median_per_mut])
dt[, mutants:= factor(mutants, levels=unique(dt[order(median_per_mut), mutants]))]

# assign color by relation to WT
dt[, rel_to_wt := ifelse(median_per_mut < wt_med, 'Smaller than WT', 'Larger than WT'),
   by = mutants]
dt[mutants == 'WT', rel_to_wt := 'WT']

p <- ggplot(dt, aes(mutants, pro_uptake, fill = rel_to_wt)) +
  geom_boxplot() +
  geom_jitter(width = 0.4) +
  labs(y = "Proline Uptake")

# ggplotly(p)
p
```



```r
# Another solution with bar plot:
summary_dt <- dt[, .(mean = mean(pro_uptake),
                     sd = sd(pro_uptake)),
                 by = "mutants"]
x_order <- summary_dt[order(mean), mutants]
summary_dt[, mutants := factor(mutants, levels = x_order)]
dt[, mutants := factor(mutants, levels = x_order)]
# get wt mean
wt <- summary_dt[mutants == "WT", mean]
# group mutants to larger and smaller than wt
summary_dt[, color := ifelse(mean > wt, "Larger",
                             ifelse(mean == wt, "WT", "smaller"))]

ggplot(summary_dt) +
  geom_bar(aes(mutants, mean, fill = color), stat='identity') +
  geom_errorbar(aes(mutants, ymax=mean+sd, ymin=mean-sd), width = 0.2) +
  geom_jitter(data = dt, aes(mutants, pro_uptake))
```



Section 5 - Case Study Feedback

Feedback for the report (Rmd) with the entire analysis

- ☐ Does the notebook run and create all figures from the presentation?
- ☐ Is the notebook cleaned-up?
- ☐ Is the report (Rmd) stand alone? (Only Rmd needed to understand the performed analysis. It should contain explanations/interpretations and code.)
- ☐ Is the data after the data preparation tidy? Is the definition of an observation clear?

Feedback on the presentation / slides

General considerations

- ☐ Has the presentation a clear structure?
- ☐ Did the presentation tell a clear and convincing story?
- ☐ Did the presenter stick to the 7 min limit?

Introduction

- ☐ Did the presentation start with a short motivation?
- ☐ Are the goals of the analyses formulated at the beginning of the presentation?

Data Preparation

- ☐ Were important data preparation steps explained during the presentation?
- ☐ Was additional data used in the presentation? If, was it made clear how that data was obtained?

Data Exploration/Analysis

- ☐ Were the stated claims communicated well?
- ☐ Were all stated claims supported by appropriate figures?
  - Was the appropriate plot type (line plot, scatter plot, box plot, violin plot) selected for the data shown?
  - Did the plot support the claim? Is there a better alternative to visualize the claim?
  - Was an associative plot used to show the relationship stated in the claim.
- ☐ Did the figures follow the plotting guidelines (no double-encoding, good paper-ink-ratio)?
- ☐ Were alternative interpretations of the associations discussed (directionality, effect of a third variable, robustness)?
- ☐ Where there any circularities in the presented claims?

Conclusion

- ☐ Did the presentation end with a conclusion slide recapping the main findings?
- ☐ Did the claims answer the problem/goals formulated in the motivation?
- ☐ Did the presentation help to learn something that we/you did not know before? (non-evident claims/association/pattern/relationship in the data presented)